SAND 2003-3124P

ASCI Red Storm and and Supercomputer Scalability

Dr. Erik P. DeBenedictis
Sandia National Laboratories

Symposium on Supercomputations
Sarov, Russia

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Outline

- Red Storm Overview
- Scalability
- Light Weight Kernel

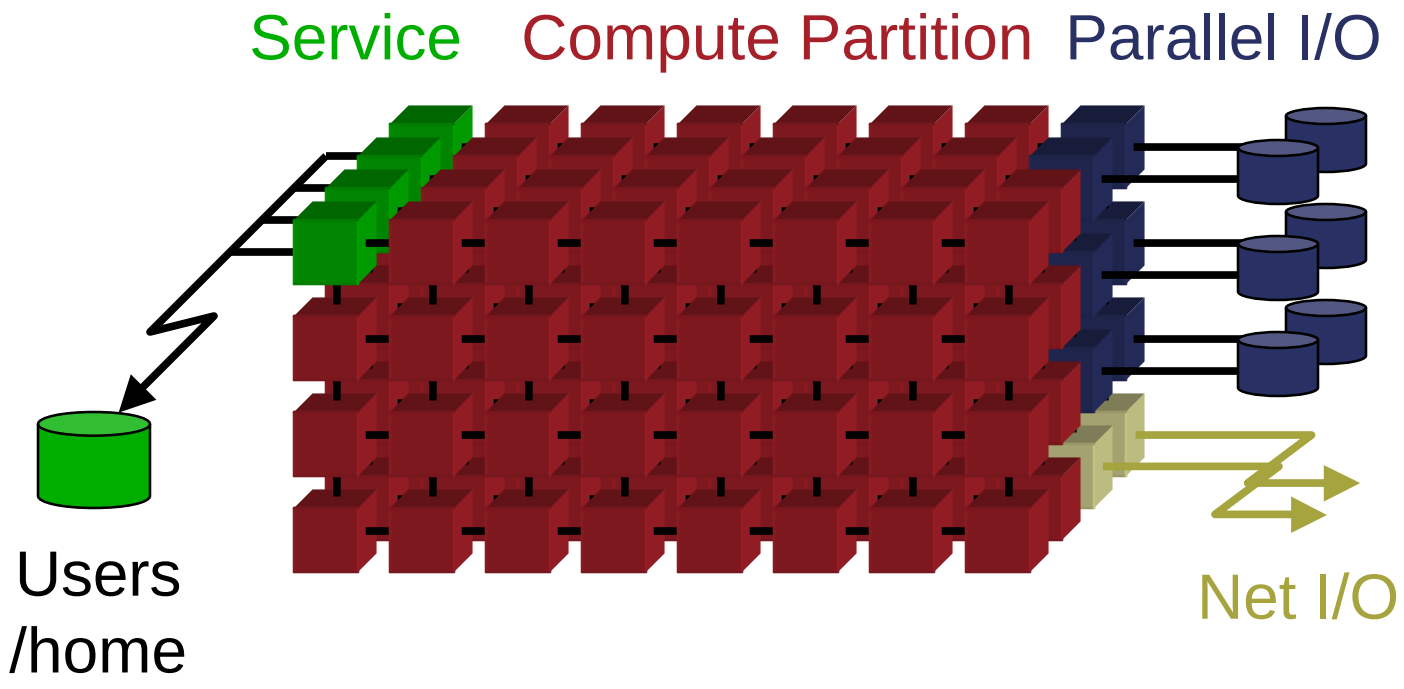
Sandia National Laboratories

Project Overview

- Red Storm is a nominally 40 TFlops supercomputer that is part of the Advanced Simulation and Computation (ASCI) program
- Red Storm was specified by and is being procured by Sandia National Laboratories
- Red Storm is being manufactured by Cray, Inc.
- Initial delivery to Sandia is scheduled for May, 2004

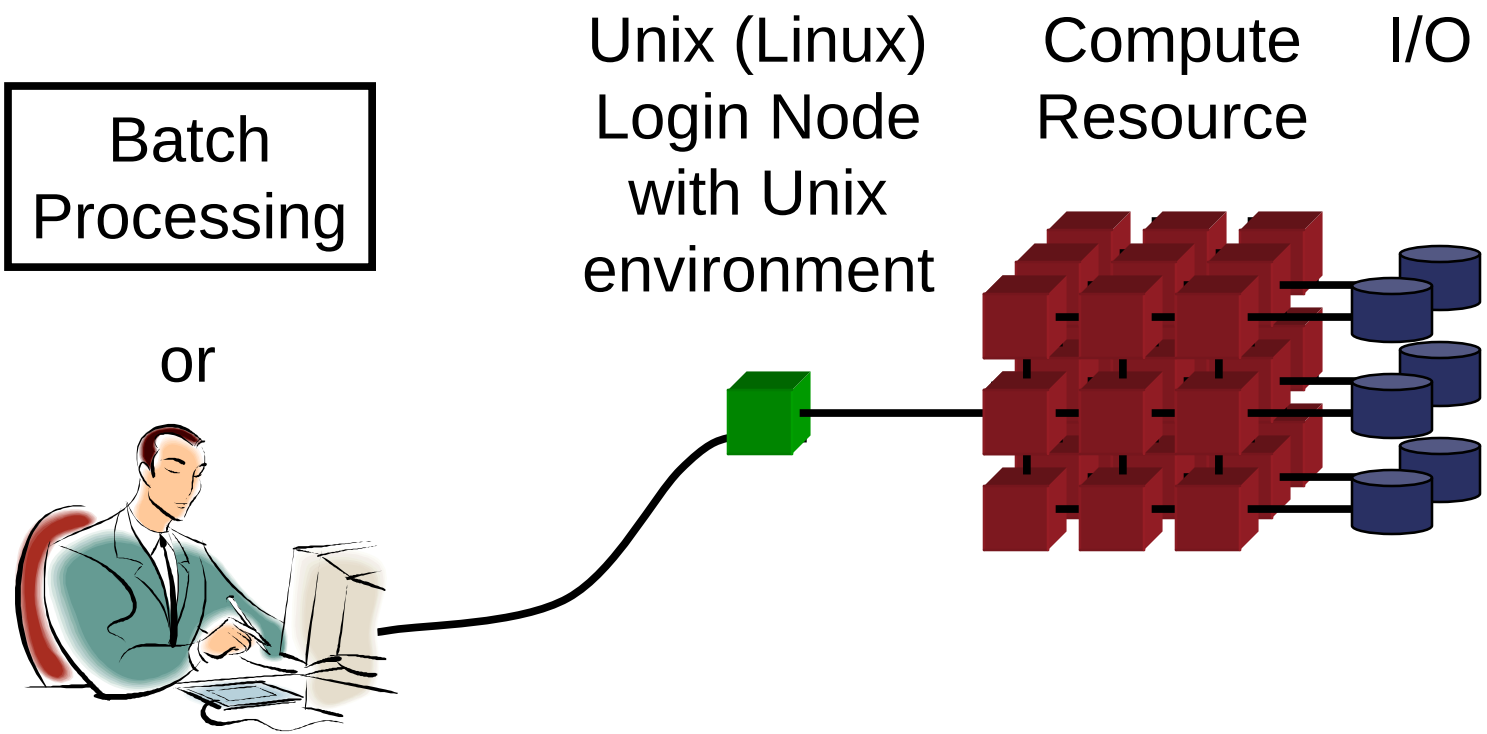
Sandia National Laboratories

Red Storm is a Massively Parallel Processor

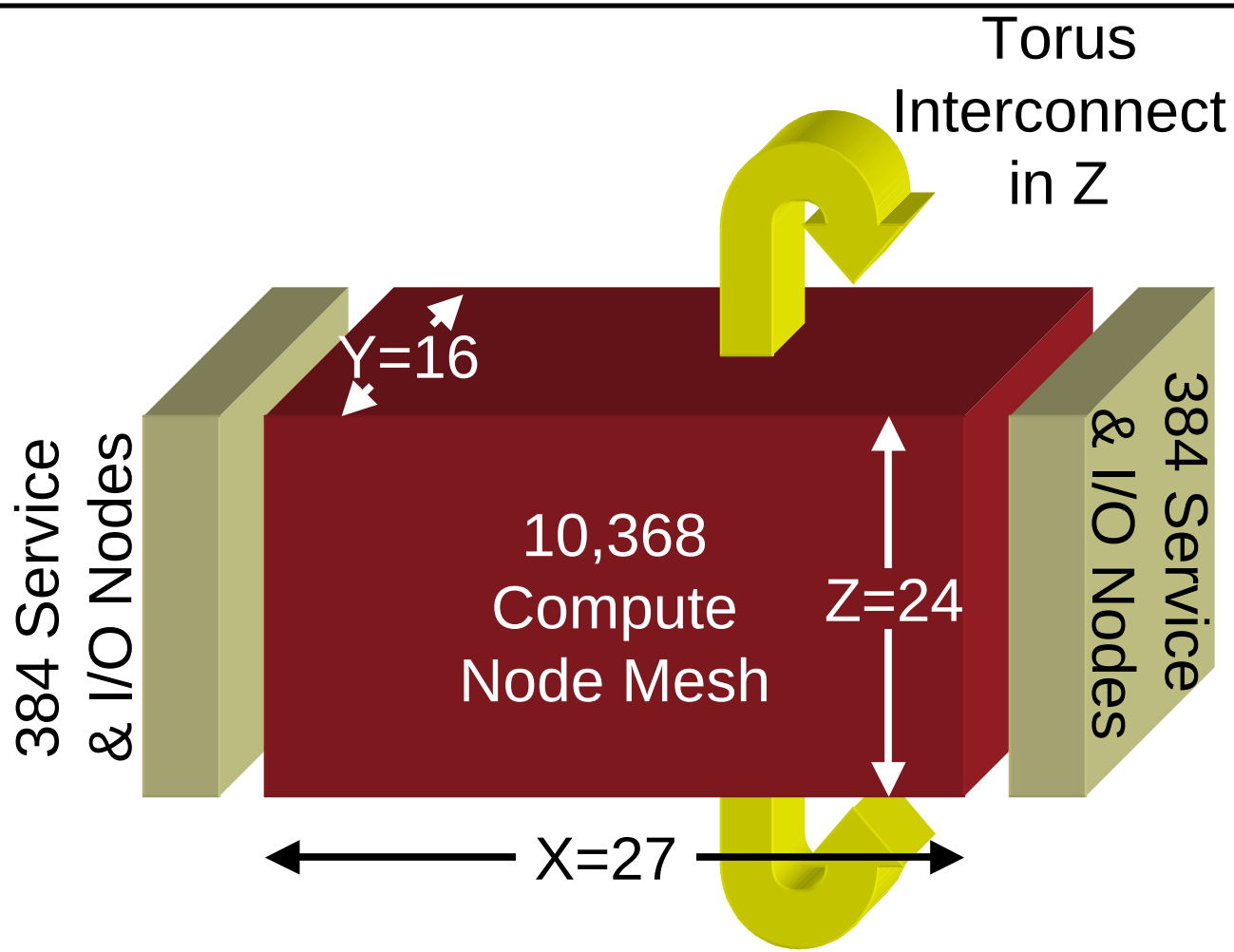


Sandia National Laboratories

Usage Model


Batch Processing
or
Unix (Linux) Login Node with Unix environment
Compute Resource
I/O

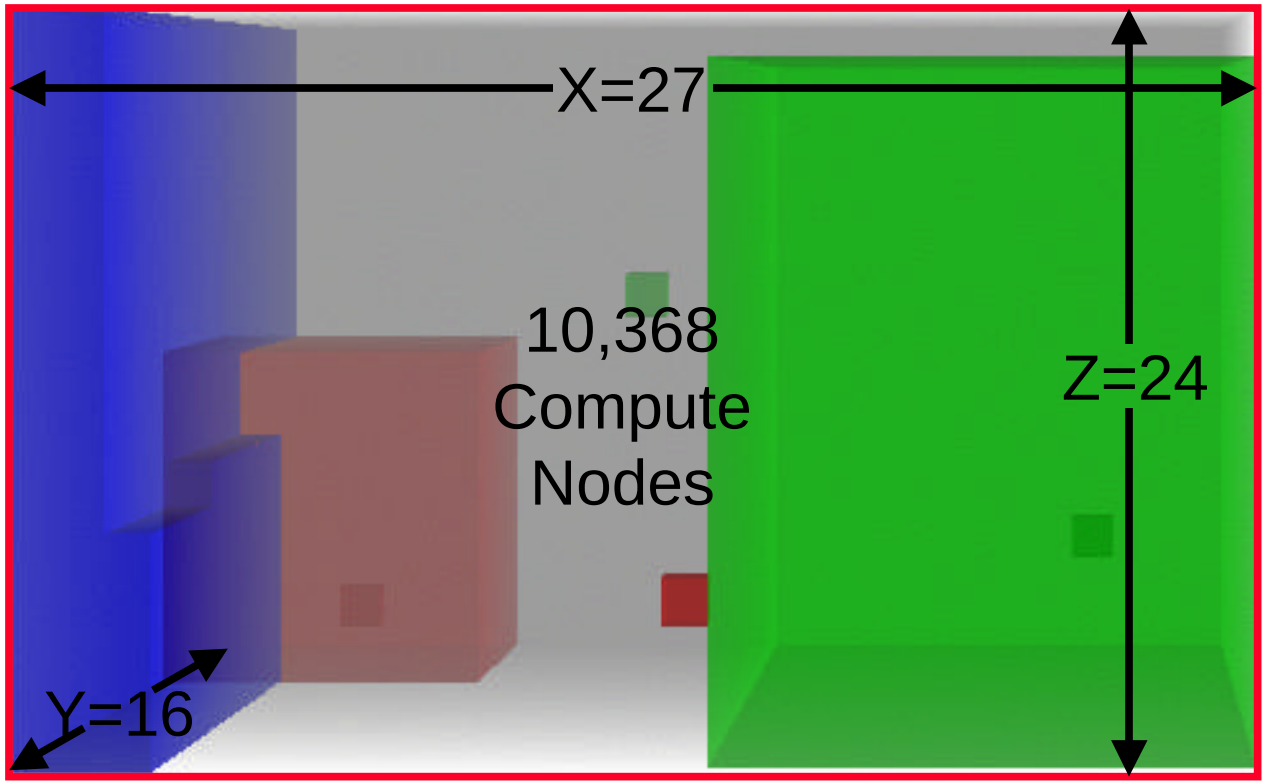
27×16×24 3D Mesh/Torus + I/O



Sandia National Laboratories

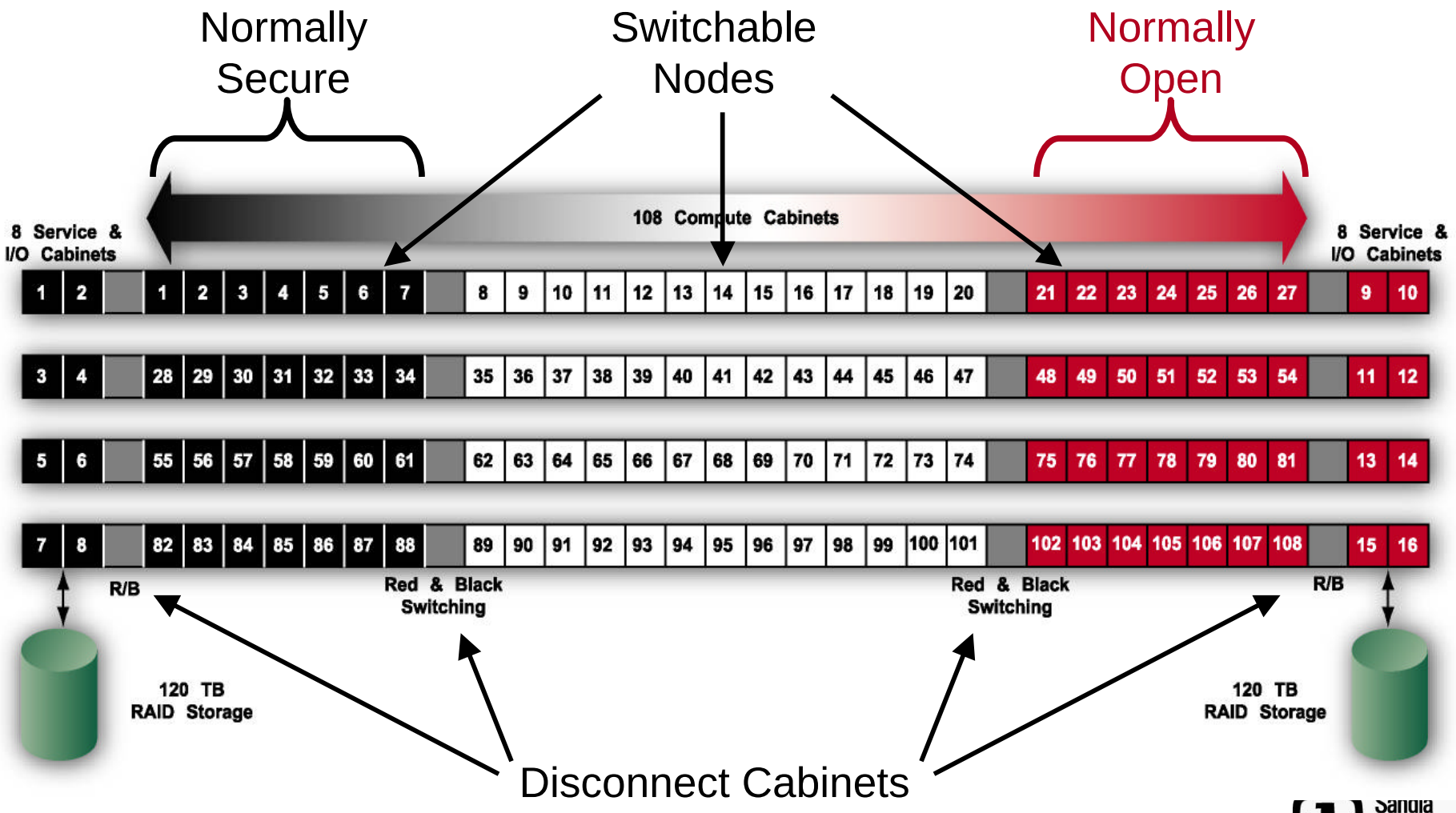
Space Sharing of Jobs

- **Jobs occupy disjoint regions simultaneously**
- **Example – red, green, and blue jobs:**



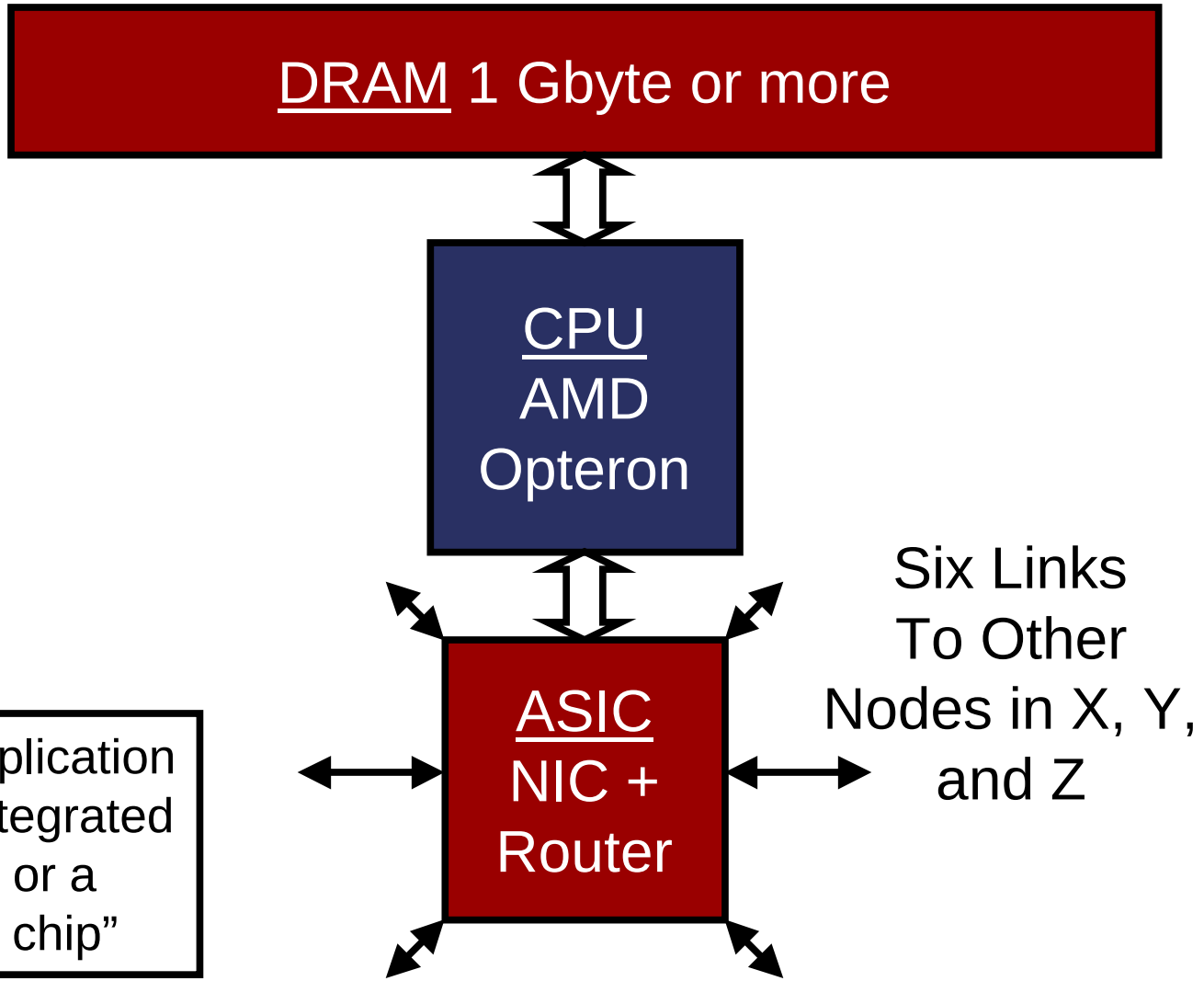
Sandia National Laboratories

Red Storm Hardware Overview



Sandia National Laboratories

Node Architecture



ASIC = Application Specific Integrated Circuit, or a "custom chip"

Sandia National Laboratories

Scalability

- Communications is the key concern
  - Amdahl's Law limits the scalability of parallel computation…
  - but not due to serial work in the application

- Why?

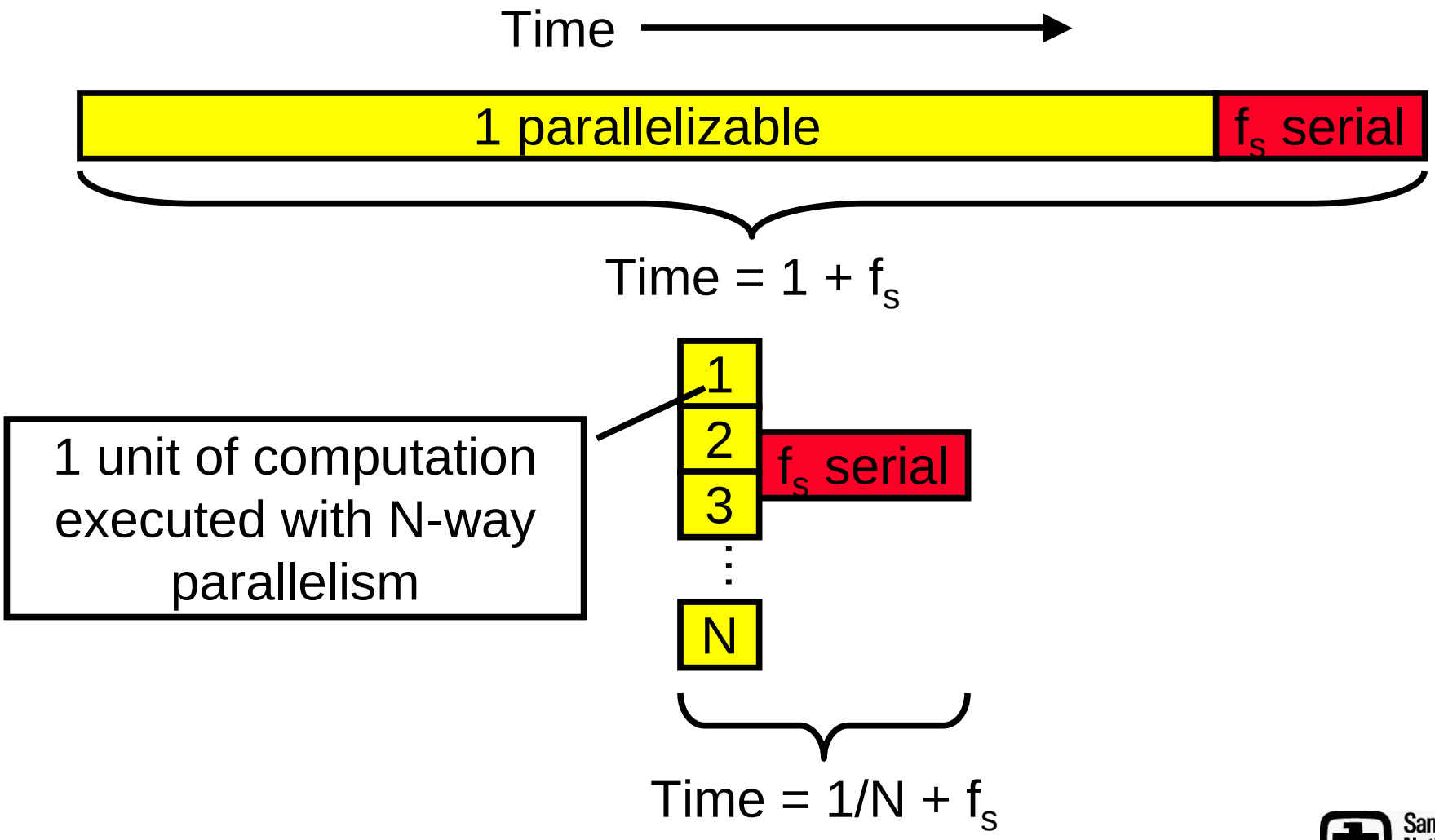
Amdahl's Law

$$S_{Amdahl}(N) = [1 + f_s]/[1/N + f_s]$$

where S is the speedup on N processors and f_s is the serial (non-parallelizable) fraction of the work to be done.

Amdahl says that in the limit of an infinite number of processors, S cannot exceed $[1+ f_s]/ f_s$. So, for example if $f_s = 0.01$, S cannot be greater than 101 no matter how many processors are used.

Amdahl's Law Picture



Sandia National Laboratories

Amdahl's Law

Example:

How big can f_s be if we want to achieve a speedup pf 8,000 on 10,000 processors (80% parallel efficiency)?

Answer:

f_s must be less than 0.000025 !

Amdahl's Law

Contrary to Amdahl & most folks' early expectations, well designed codes on balanced systems can routinely do this well or better!

However in applying Amdahl's Law, we neglected the overhead due to communications.

A Realistic View of Amdahl's Law

The actual scaled speedup is more like

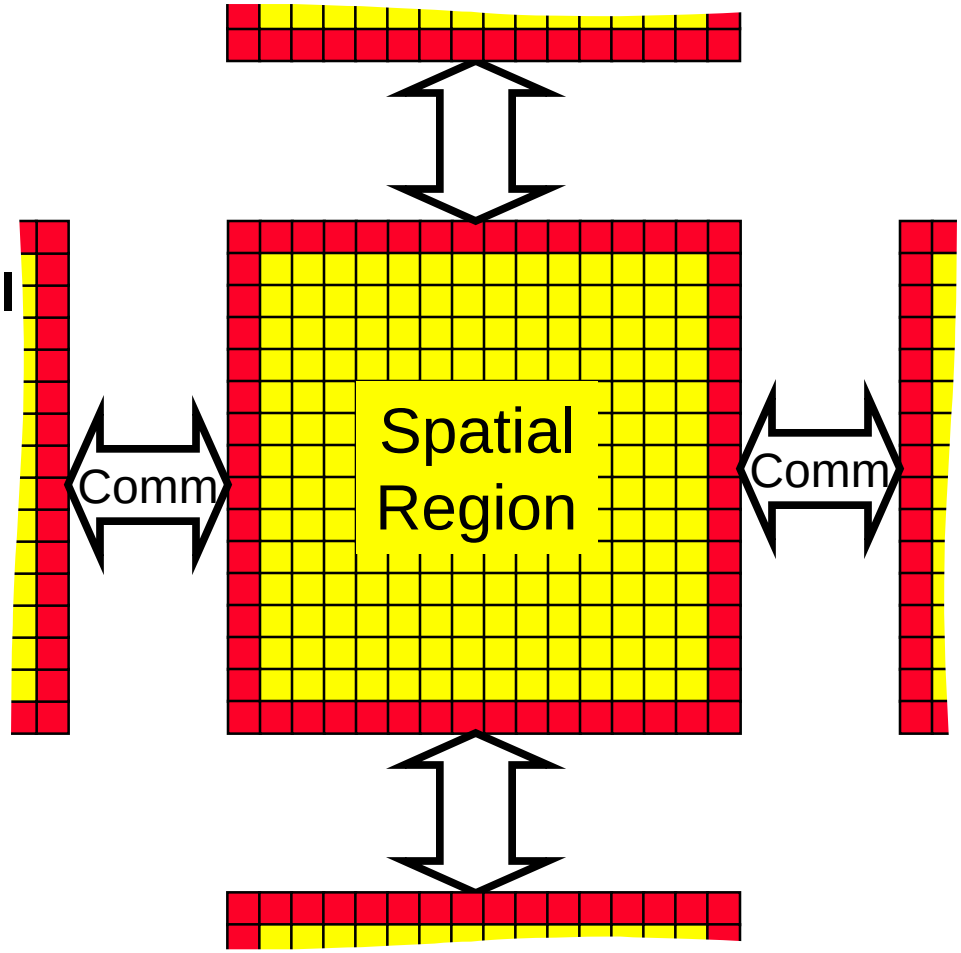
$$S(N) \sim S_{Amdahl}(N)/[1 + f_{comm} \times R_{p/c}],$$

where f_{comm} is the fraction of work devoted to communications and $R_{p/c}$ is the ratio of processor speed to communications speed.

Realistic Picture of Amdahl's Law

- **Problem is a physical simulation in two dimensions**
- **Ratio of boundary (■) to all points (□+■) is f_{comm}**
- **Boundary runs at slower due to communications, say ratio of $R_{p/c}$**
- **Communications will slow execution by factor of**

$$\frac{1}{1 + f_{comm} \times R_{p/c}}$$



Sandia National Laboratories

Implications of Realistic Amdahl's Law

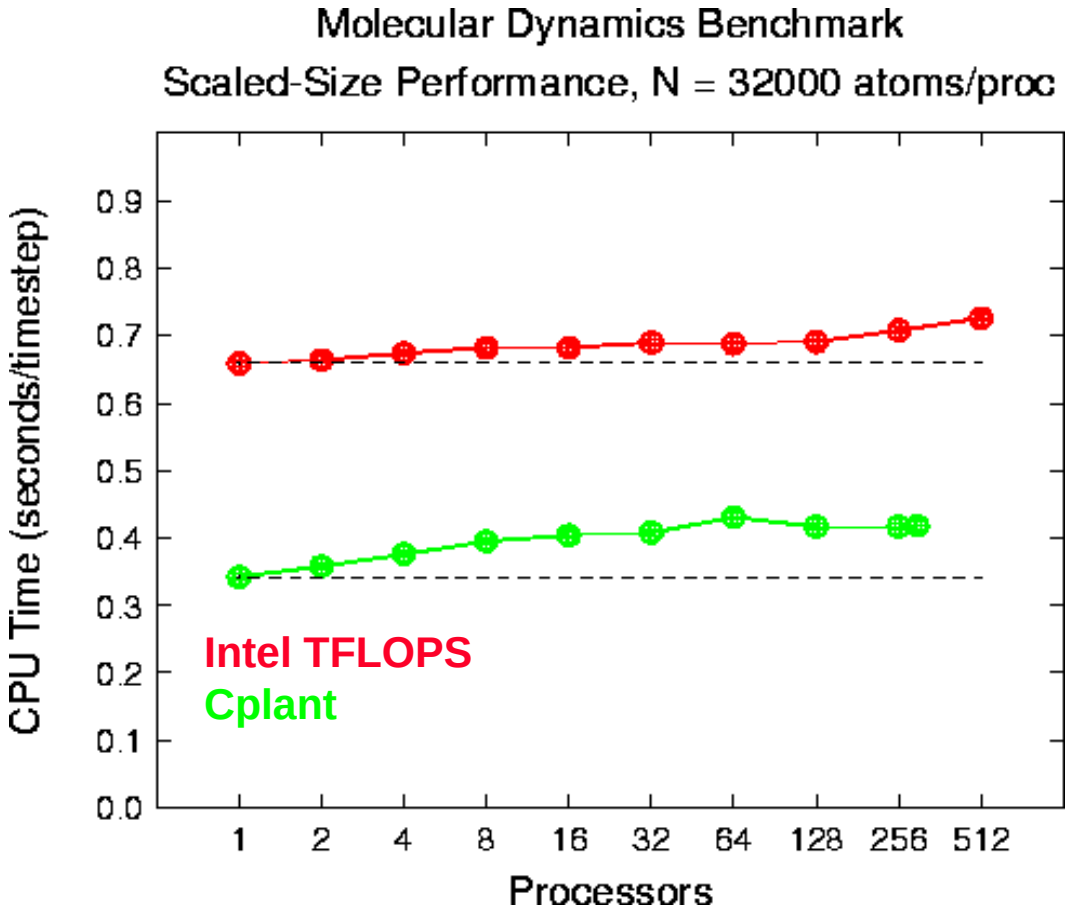
- Let's consider three cases on two computers:
  - The two computers are identical except that one has
    - $R_{p/c}$ = 1 Byte/FLOP (fast communications)
    - $R_{p/c}$ = 0.05 Byte/FLOP (not so fast communications)
  - The three cases are
    - f_{comm} = 0.01,
    - f_{comm} = 0.05, and
    - f_{comm} = 0.10

Real Amdahl's Law Efficiency

Efficiency	F_{comm} = .01 99% comp. dominated	F_{comm} = .05 95% comp. dominated	F_{comm} = .1 90% comp. dominated
$R_{p/c}$ = 1 Time to send a number ≈ time for an op on it	99% Efficient	95% Efficient	90% Efficient
$R_{p/c}$ = 0.05 Time to send a number ≈ time for 20 ops on it	83% Efficient	50% Efficient	33% Efficient

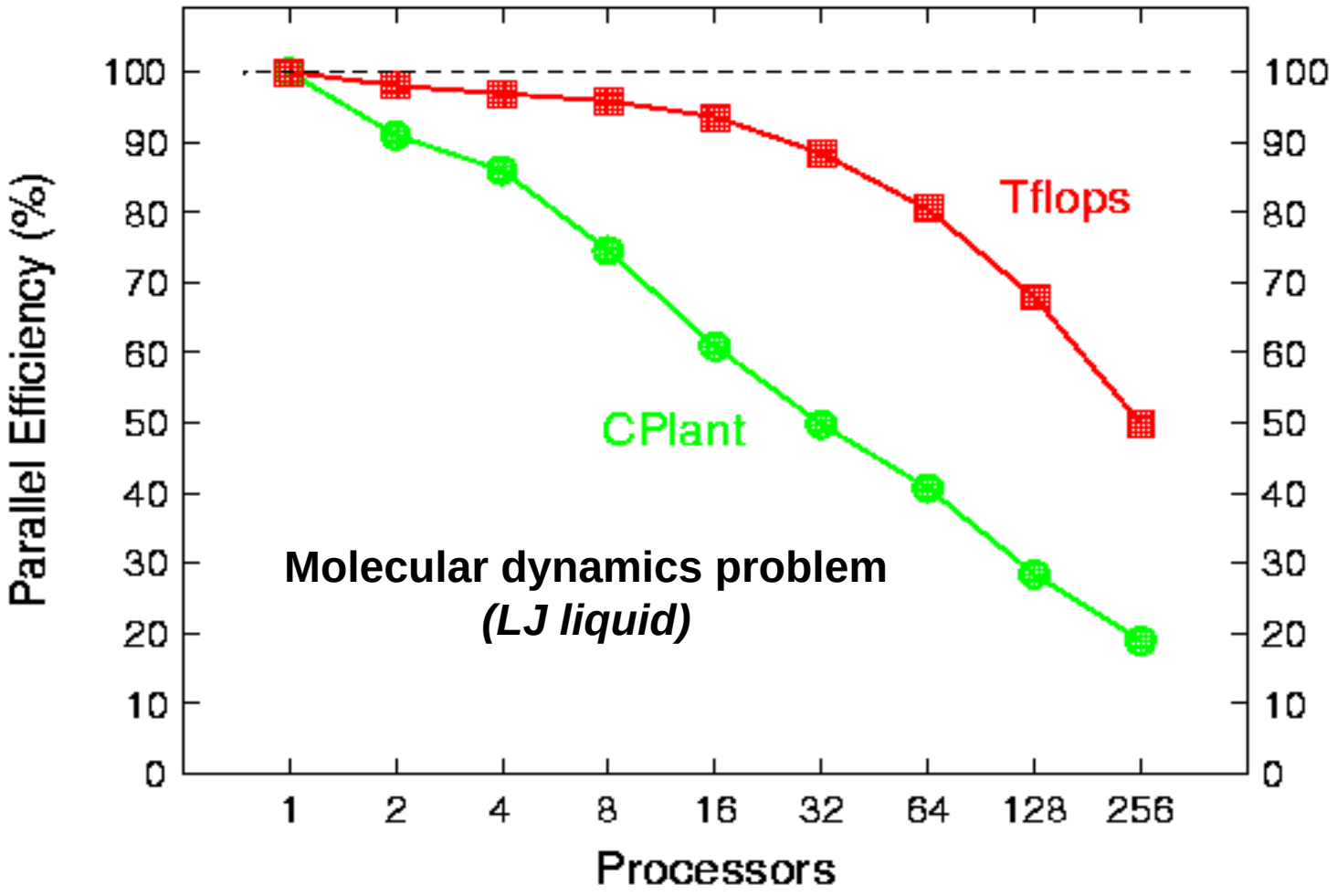
Sandia National Laboratories

Sandia Experience with $R_{p/c}$



Sandia National Laboratories

Sandia Experience with $R_{p/c}$



Sandia National Laboratories

Importance of Balanced Communications

- **A "well-balanced" architecture is nearly insensitive to communications overhead**
- **By contrast a system with weak communications can lose over half its power for applications in which communications is important**
- **Red Storm has been designed with $R_{p/c} \approx 1$**

Sandia National Laboratories

Comparisons of Communications Balance

Machine	Node Speed Rating(MFlops)	Link BW (Mbytes/s)	Ratio (Bytes/flop)
ASCI RED	**400**	**800(533)**	**2(1.33)**
T3E	**1200**	**1200**	**1**
ASCI RED**	**666**	**800(533)**	**(1.2)0.67**
Cplant	**1000**	**140**	**0.14**
Blue Mtn*	**500**	**800**	**1.6**
BlueMtn**	**64000**	**1200 (9600*)**	**0.02 (0.16*)**
Blue Pacific	**2650**	**300 (132)**	**0.11 (0.05)**
White	**24000**	**2000**	**0.083**
Q*	**2500**	**650**	**0.2**
Q**	**10000**	**400**	**0.04**

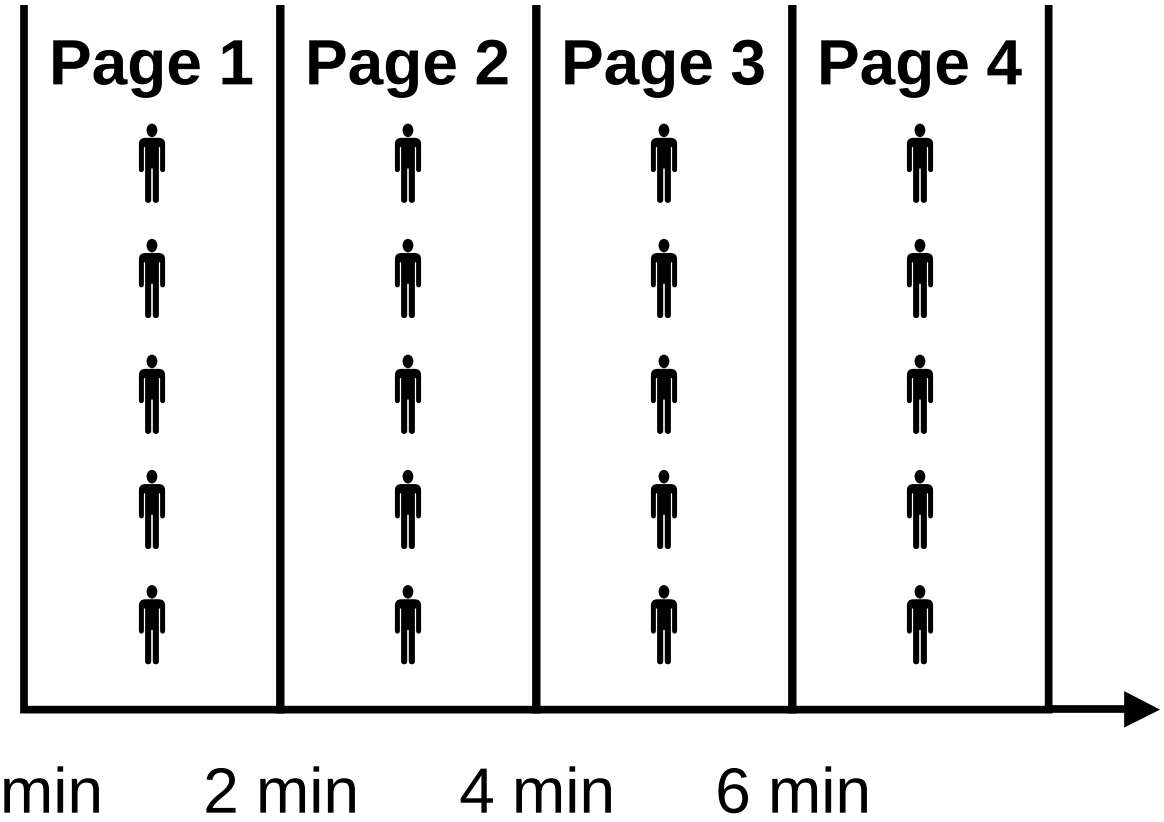
Sandia National Laboratories

Light Weight Kernel

- Sandia has had very good experiences with LWK
  - Sandia-University of New Mexico Operating System (SUNMOS)
  - Cougar
  - Puma
  - Now Catamount (tell story about name)
- Why?
  - Timing stability
  - Maturity

LWK & Musical Rehearsal

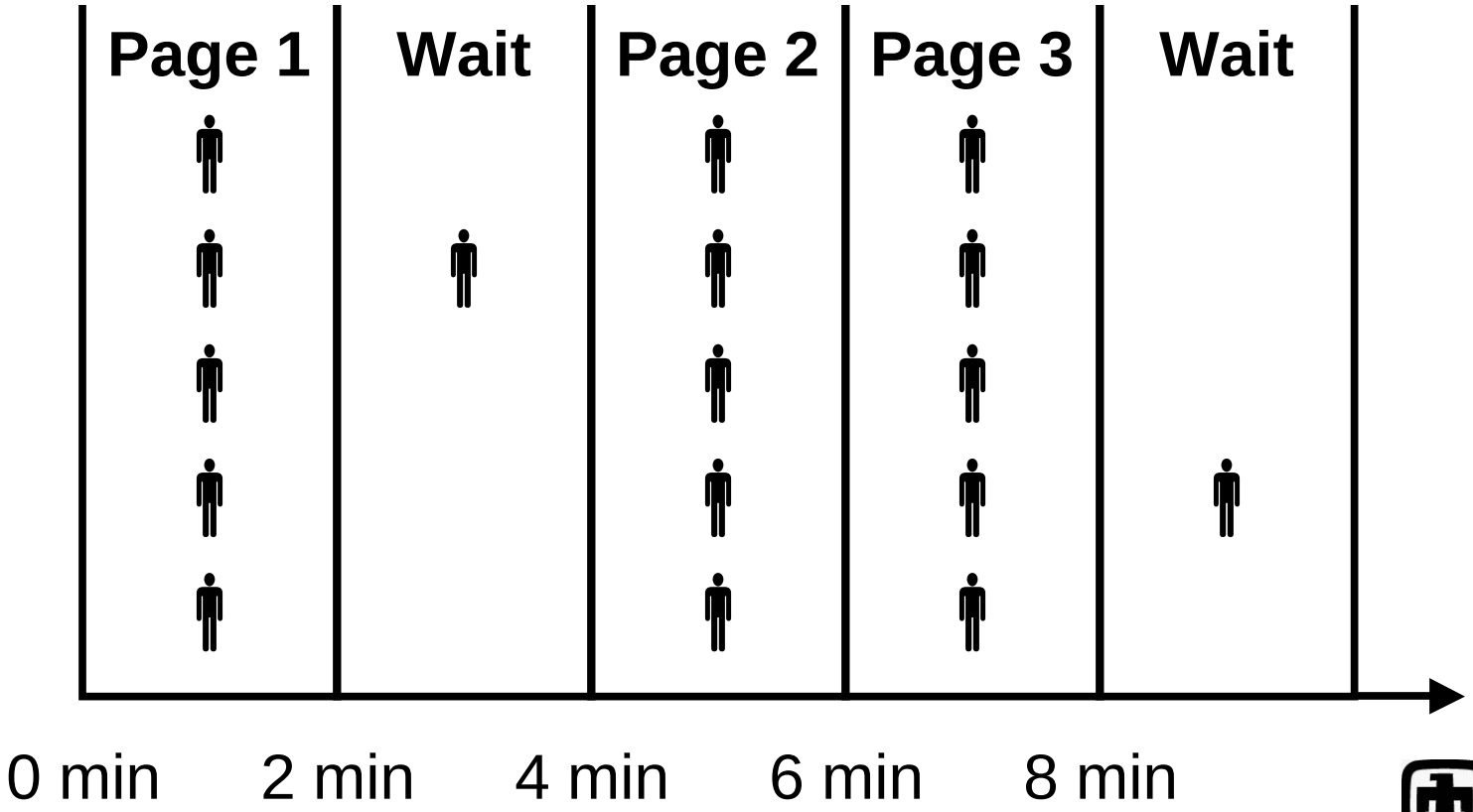
- **N musicians Rehearsing 2 Minute Pages of Music**



Sandia National Laboratories

Musical Rehearsal with Breaks

- 2 Minute Pieces with Asynchronous Breaks



Sandia National Laboratories

Breaks in MPP Systems Software

- Unix, Linux, any OS
  - Kernel memory allocation
  - TCP/IP backoff calculations
  - Routing tables
  - Clock synchronization
  - Scheduler
  - Etc., full list unknown, but has been extremely problematic with DOE labs
- Light Weight Kernel
  - None

Run Time Impact of Unix Systems Services

- Say breaks take 50 μS and occur once per second
  - On one CPU, wasted time is 50 μs every second
    - Negligible .005% impact
  - On 100 CPUs, wasted time is 5 ms every second
    - Negligible .5% impact
  - On 10,000 CPUs, wasted time is 500 ms
    - Significant 50% impact
- Red Storm will have 10,000 CPUs, hence LWK approach important

Conclusions

- Red Storm is under construction as a 40 TFLOPS supercomputer
  - Delivery in about one year
- Built on engineering principles of ASCI Red
  - Expected to perform 7x as efficiently
- Performance analysis indicates that the architecture can be scaled considerably beyond Red Storm